ContextFace: Generating Facial Expressions from Emotional Contexts

Supplementary Material

1. Prompts for Context Annotation

system prompt

Without adding any additional explanation, please generate sentences following the format I requested. Even if the person in the photo is a famous character from movies or dramas, please create a context without using knowledge about them. Please avoid directly mentioning the emotion labels provided in the prompt.

user prompt

<image> The emotion label of the person in the red box is {emotion}. Based on this photo, please create one sentence each imagining the specific situation this person is in and what this person might have said in following format:
[Expected situation]: One sentence
[Expected quote]: One sentence

CAER-S and SFEW datasets contain images extracted from movies and TV shows. To prevent the LLM from generating unrealistic situations and quotes based on its prior knowledge of famous movies or TV shows, we implemented specific constraints in the system prompt to ensure the generation of more naturalistic and everyday scenarios. In the user prompt, we include both the image and its corresponding emotion category. Since CAER-S and SFEW contain natural movie and TV scenes where facial expressions may not clearly convey the intended emotion, providing the explicit emotion label helps ensure more accurate contextual generation.

2. Dataset Quality Assessment

To assess the quality of our proposed dataset, we conducted an evaluation using GPT-4.1, a strong large language model. We randomly sampled 1,000 instances from the CAER-S-C dataset used for training and evaluated them across multiple quality dimensions, as shown in Tab. 1. The evaluation assessed four key quality metrics:

Quality Metric	**Score (0-5)**
Emotional Alignment	4.83
Quote-Situation Coherence	4.98
Scenario Realism	4.96
Overall Quality	4.93
Samples Sufficient for Training	98.4%
Samples with Ambiguity	0.5%

Table 1. Quality assessment results using GPT-4.1 evaluation

- **Emotional Alignment** (4.83/5): The degree to which the emotional labels accurately reflect the emotions expressed in the context
- **Quote-Situation Coherence** (4.98/5): The consistency and logical connection between the quoted text and the situational context
- **Scenario Realism** (4.96/5): The extent to which the scenarios and quotes reflect realistic, everyday situations
- **Overall Quality** (4.93/5): A holistic assessment of the dataset's quality

Additionally, we evaluated the dataset's coverage and completeness, finding that 98.4% of samples were sufficient for training purposes, while only 0.5% were considered suboptimal for training.

3. Ablation Study

Loss	$L_2 \downarrow$	$FD \downarrow$
L_1(Situation)	0.30	25.16
L_2(Situation)	**0.07**	**0.81**
L_1(Quote)	0.28	26.01
L_2(Quote)	**0.10**	**1.95**

Table 2. Comparison of L1 and L2 loss function

Tab. 2 presents a comparison of different loss functions for facial expression coefficient optimization, demonstrating the superior performance of L2 loss. This can be attributed to the inherent mathematical consistency: expression parameters in FLAME [3] are normalized by their standard deviations, which constitutes an L2-norm statistic. Therefore, L2 loss provides an optimization framework that is naturally aligned with the data distribution.

4. Additional Experiments

4.1. Evaluation on MER2023 Dataset

Tab. 3 shows evaluation results on a subset of the MERR2023 [4] dataset. We selected this dataset as it provides contextual emotion data from drama scenes with clearly visible faces, making it suitable for our context-aware emotion recognition task. We converted the video-based data to static images by extracting peak frames using the indexing method proposed by Emotion-LLaMA [1]. We applied the same contextual augmentation method used for SFEW-C and CAER-S-C datasets to these extracted frames. For evaluation, we randomly selected 203 samples from the MERR dataset and generated associated quotes using Claude API to create our test dataset. We focused on four overlapping emotion categories (Angry, Sad, Happy, Surprise) between MERR and our main dataset, excluding 'Worried' due to consistent misclassification as 'Fear'. Results show that all models achieve high performance on clear emotions (Angry, Happy, Sad) when quotes are provided, owing to LLMs' inherent reasoning capabilities. However, for challenging categories like Surprise, which typically show lower accuracy in standard datasets, our model demonstrates superior performance due to enhanced emotional reasoning from comprehensive training on emotion-rich data.

4.2. Performance Comparison with Emotion-trained Models

Tab. 4 presents the performance comparison with emotion-trained models on our SFEW-C dataset. While Emotion-LLaMA and LLaVA-13b-Emotic-finetuned were also trained on contextual emotion datasets, they used different finetuning approaches from ours. ContextFace shows better performance in both contextual settings, which may be attributed to our model being specifically trained to predict emotions given detailed contextual information.

Method	**Hap**	**Sad**	**Ang**	**Sur**	**UAR**	**WAR**
with quote						
BLIP-13B [2]	86.54	81.90	89.25	0	64.28	85.45
llava-1.5-13B [5]	97.78	**94.02**	95.92	66.67	96.77	95.52
LLaVA-NEXT-13B [6]	97.78	92.44	94.85	66.67	96.27	94.54
Qwen2.5-VL-7B [7]	**98.90**	93.10	94.85	40.00	96.37	94.75
ContextFace (ours)	97.83	**94.02**	**96.41**	**100**	**97.32**	**96.13**

Table 3. Zero-shot emotion recognition performance in F1 scores on MERR Dataset. categories: Hap (Happy), Sad (Sad), Ang (Angry), Sur(Surprise). UAR: Unweighted Average Recall, WAR: Weighted Average Recall.

Method	Hap	Sad	Neu	Ang	Sur	Dis	Fea	UAR	WAR
with situation									
Emotion-LLaMA [1]	86.75	91.14	53.91	94.19	66.06	72.22	69.92	77.80	75.78
LLaVA-13b-Emotic-finetuned [8]	**96.50**	91.14	88.31	89.94	77.36	76.92	**85.71**	86.07	84.41
ContextFace (ours)	94.20	**95.89**	**92.77**	**94.94**	**85.45**	**85.71**	84.31	**90.13**	**90.68**
with quote									
Emotion-LLaMA [1]	74.47	90.07	60.00	90.24	75.93	82.05	80.43	78.94	78.59
LLaVA-13b-Emotic-finetuned [8]	92.75	**96.60**	**92.31**	90.59	80.67	80.95	82.22	87.42	86.85
ContextFace (ours)	**96.60**	92.21	84.93	**94.48**	**91.74**	**95.65**	**92.78**	**93.21**	**94.89**

Table 4. Emotion-trained models performance in F1 scores on SFEW-C Dataset. Emotion categories: Hap (Happy), Sad (Sad), Neu (Neutral), Ang (Angry), Sur (Surprise), Dis (Disgust), Fea (Fear). UAR: Unweighted Average Recall, WAR: Weighted Average Recall.

References

[1] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning, 2024. 2, 3

[2] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 3

[3] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans, 2017. 2

[4] Zheng Lian, Haiyang Sun, Licai Sun, Kang Chen, Mingyu Xu, Kexin Wang, Ke Xu, Yu He, Ying Li, Jinming Zhao, Ye Liu, Bin Liu, Jiangyan Yi, Meng Wang, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning, 2023. 2

[5] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3

[6] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 3

[7] Qwen Team. Qwen2.5-vl, 2025. 3

[8] yetesam. Llava-finetuned-contextual-emotion-recognition, 2024. Hugging Face Model Hub. 3